Metagenome Overview

MG-RAST ID 4636848.3 Download Analyze Search

Metagenome Name ssr_15372__R1__L001
PI
Organization
Visibility Private
Static Link http://metagenomics.anl.gov/linkin.cgi?metagenome=4636848.3

NCBI Project ID -
GOLD ID -
PubMed ID -

Delete | Share | Edit Name | Make Public

METAGENOME SUMMARY

Dataset ssr_15372__R1__L001 was uploaded on 06/06/2015 and contains 126,930 sequences totaling 18,297,630 basepairs with an average length of 144 bps. The piechart below breaks down the uploaded sequences into 3 distinct categories.

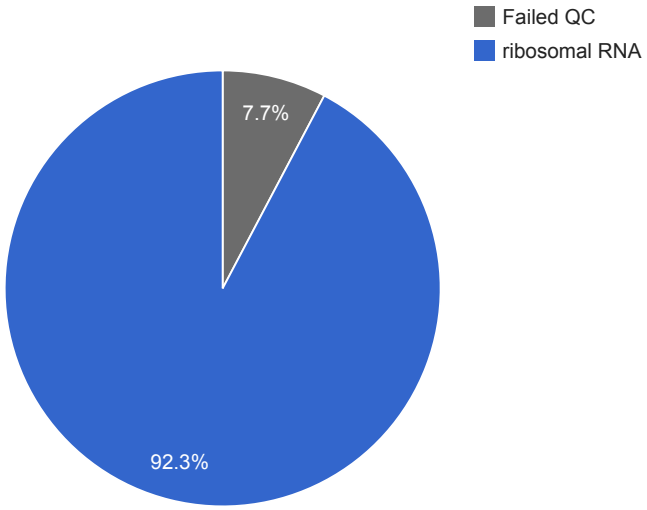
10,413 sequences (8.2%) failed to pass the QC pipeline. Of the sequences that passed QC, 124,407 sequences (98.0%) contain ribosomal RNA genes. 0 (0.0%) of the sequences that passed QC have no rRNA genes.

The analysis results shown on this page are computed by MG-RAST. Please note that authors may upload data that they have published their own analysis for, in such cases comparison within the MG-RAST framework can not be done.

TABLE OF CONTENTS

DOWNLOAD data and annotations
ANALYZE annotations in detail.
SEARCH through annotations.

Sequence Breakdown



Note: Sequences containing multiple predicted features are only counted in one category.
Currently downloading of sequences via chart slices is not enabeled.

- Organism Breakdown
  - Taxonomic Distribution
  - Rank Abundance Plot
  - Rarefaction Curve
  - Alpha Diversity
- Technical Data
  - Statistics
  - Metadata
  - Source Distribution
  - Sequence Length Histogram
  - Sequence GC Distribution

PROJECT INFORMATION

This dataset is part of project sprague-april2015.

There are 4 other metagenomes in this project

GSC MIXS INFO

Investigation Type	**mimarks-survey**
Project Name	**sprague-april2015**
Latitude and	**-, -**

» find metagenomes within this project
» find metagenomes within this biome
» find metagenomes within this country
» find metagenomes within 10 | 30 | 100 kilometers

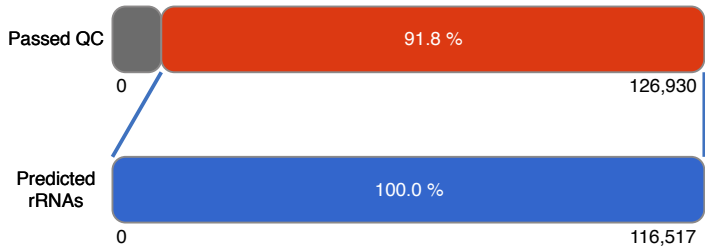
Longitude	
Country and/or Sea, Location	-
Collection Date	-
Environment (Biome)	-
Environment (Feature)	-
Environment (Material)	-
Environmental Package	-
Sequencing Method	**illumina**
More Metadata	

Analysis Flowchart

10,413 sequences failed quality control. Of the 116,517 sequences (totaling 17,008,921 bps) that passed quality control, 124,407 (106.8%) produced a total of 1,665 identified ribosomal RNAs.

Analysis Statistics

Upload: bp Count	**18,297,630 bp**
Upload: Sequences Count	**126,930**



Upload: Mean Sequence Length	**144 ± 27 bp**
Upload: Mean GC percent	**54 ± 4 %**
Post QC: bp Count	**17,008,921 bp**
Post QC: Sequences Count	**116,517**
Post QC: Mean Sequence Length	**145 ± 25 bp**
Post QC: Mean GC percent	**53 ± 4 %**
Processed: Predicted rRNA Features	**2,301**
Alignment: Identified rRNA Features	**1,665**

DRISEE [?]

Duplicate Read Inferred Sequencing Error Estimation (Keegan et al., PLoS Computational Biology, 2012)

DRISEE could not produce a profile, this is an Amplicon dataset.

DRISEE is a tool that utilizes artificial duplicate reads (ADRs) to provide a platform independent assessment of sequencing error in metagenomic (or genomic) sequencing data. DRISEE is designed to consider shotgun data. Currently, it is not appropriate for amplicon data.

Note that DRISEE is designed to examine sequencing error in raw whole genome shotgun sequence data. It assumes that adapter and/or barcode sequences have been removed, but that the sequence data have not been modified in any additional way. (e.g.) Assembly or merging, QC based triage or trimming will both reduce DRISEE's ability to provide an accurate assessment of error by removing error before it is analyzed.

KMER PROFILES [?] HIDE

Redraw the below plot using the following kmer-plot type: kmer rank abundance 15-mer

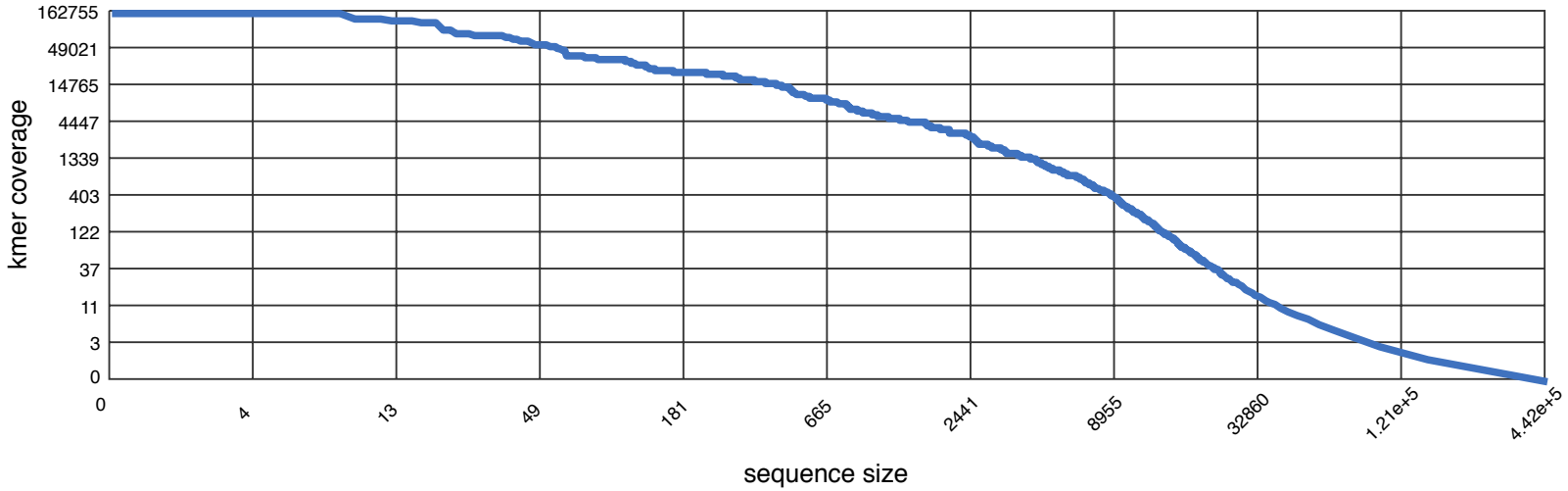
The kmer abundance spectra are tools to summarize the redundancy (repetitiveness) of sequence datasets by counting the number of occurrences of 15 and 6 bp sequences.

The kmer spectrum plots the number of distinct N-bp sequences as a function of coverage level, placing low-coverage (rare) sequences at left and high-coverage,

repetitive sequences at right. The kmer rank abundance graph plots the kmer coverage as a function of abundance rank, with the most abundant sequences at left. The ranked kmer consumed graph shows the fraction of the dataset that is explained by the most abundant kmers, as a function of the number of kmers used.

Download chart data

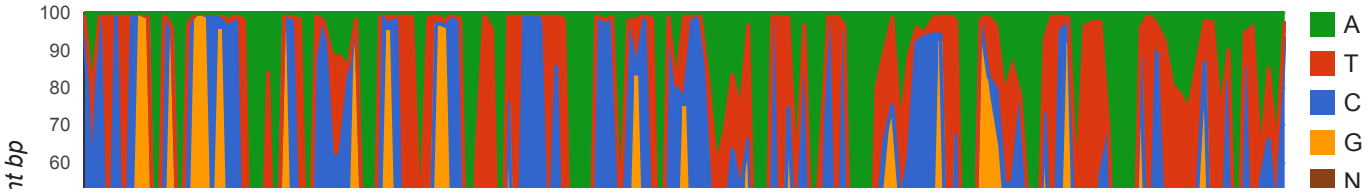
The image is currently dynamic. To be able to right-click/save the image, please click the static button static



NUCLEOTIDE POSITION HISTOGRAM [?] HIDE

These graphs show the fraction of base pairs of each type (A, C, G, T, or ambiguous base "N") at each position starting from the beginning of each read up to the first 151 base pairs. Amplicon datasets should show consensus sequences; shotgun datasets should have roughly equal proportions of basecalls.

Download chart data



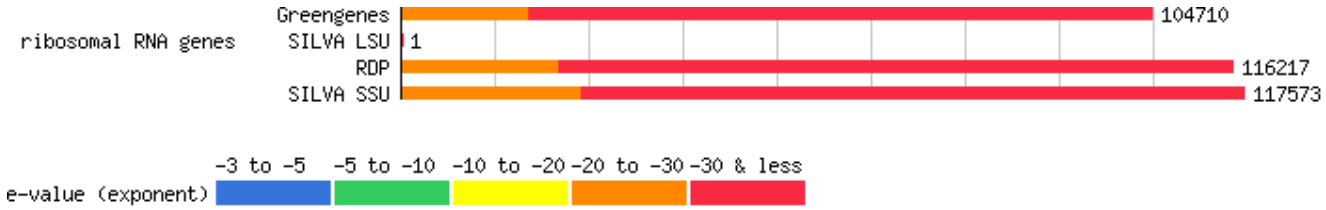
SOURCE HITS DISTRIBUTION [?] HIDE

124,407 (98.0%) of reads had similarity to ribosomal RNA genes.

The graph below displays the number of features in this dataset that were annotated by the different databases below. These include protein databases, protein databases with functional hierarchy information, and ribosomal RNA databases. The bars representing annotated reads are colored by e-value range. Different databases have different numbers of hits, but can also have different types of annotation data.

There are 15,945,780 sequences in the M5NR protein database and 309,342 sequences in the M5RNA ribosomal database. The M5NR protein database contains all the unique sequences from the below protein databases and the M5RNA ribosomal database contains all the unique sequences from the below ribosomal RNA databases.

Download chart data

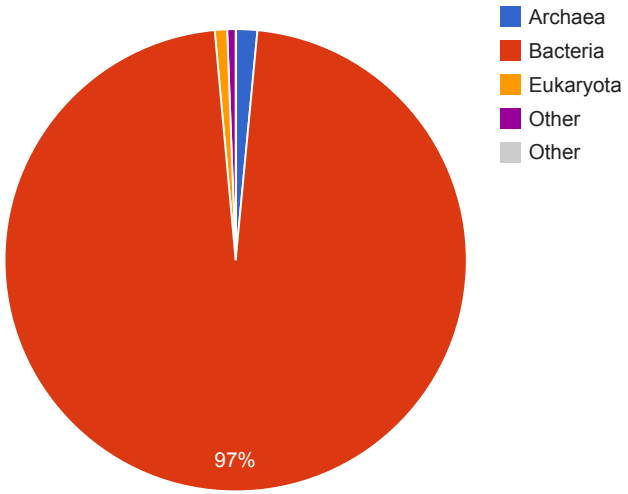


TAXONOMIC HITS DISTRIBUTION HIDE

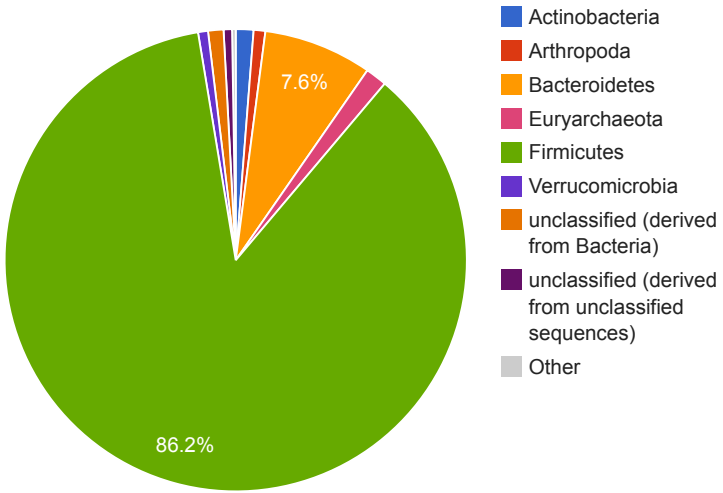
The pie charts below illustrate the distribution of taxonomic domains, phyla, and orders for the annotations. Each slice indicates the percentage of reads with predicted proteins and ribosomal RNA genes annotated to the indicated taxonomic level. This information is based on all the annotation source databases used by MG-RAST. An interactive Krona chart of the full taxonomy is also available
Click on a slice or legend to view all sequences annotated with the indicated taxonomic level in the analysis page.

View taxonomic interactive chart

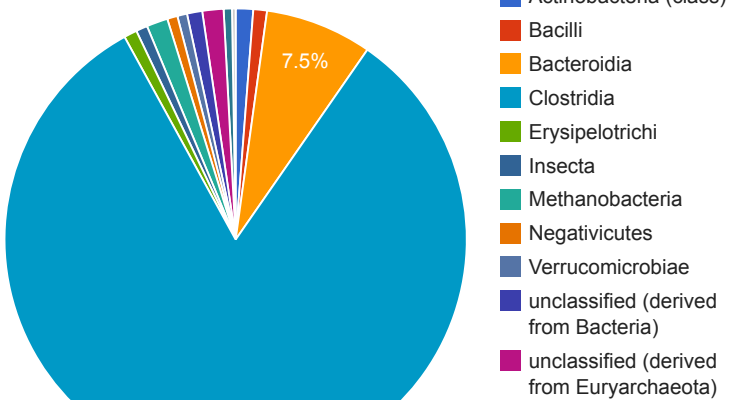
domain Download chart data



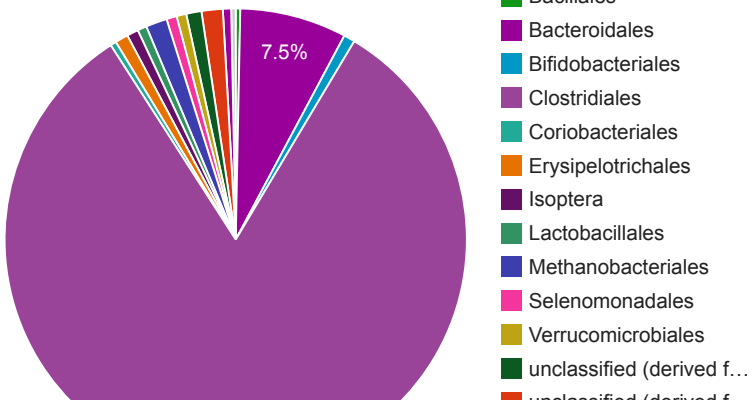
phylum Download chart data



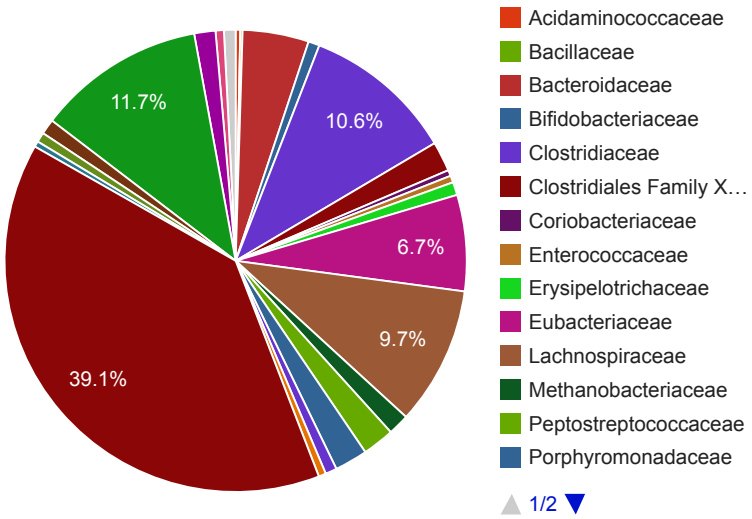
class Download chart data



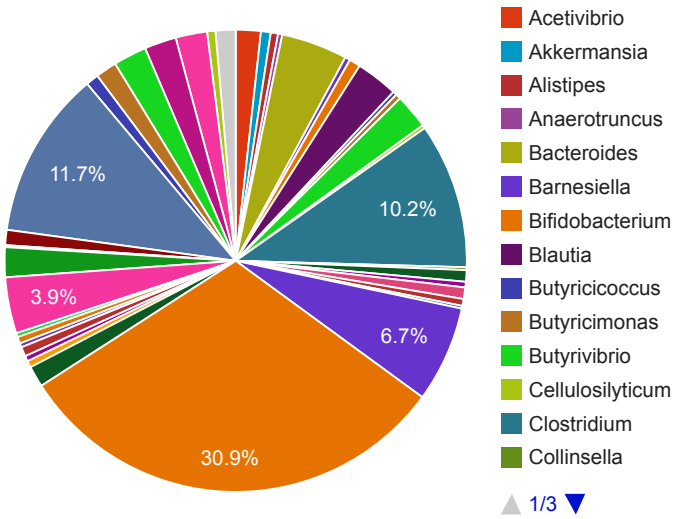
order Download chart data



family Download chart data



genus Download chart data



RANK ABUNDANCE PLOT HIDE

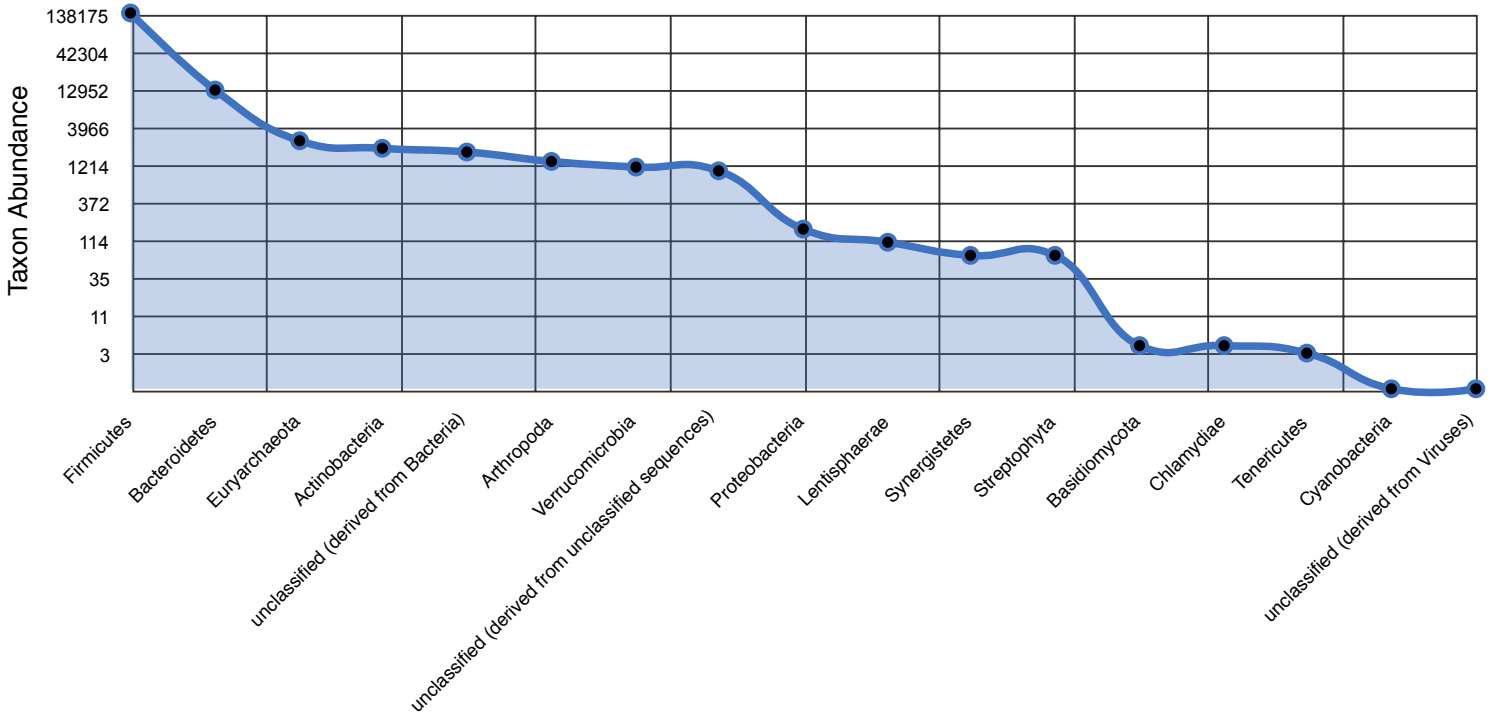
Redraw the below plot using the following taxonomic level: phylum

The plot below shows the phylum abundances ordered from the most abundant to least abundant. Only the top 50 most abundant are shown. The y-axis plots the abundances of annotations in each phylum on a log scale.

The rank abundance curve is a tool for visually representing taxonomic richness and evenness.

Download chart data

The image is currently dynamic. To be able to right-click/save the image, please click the static button static



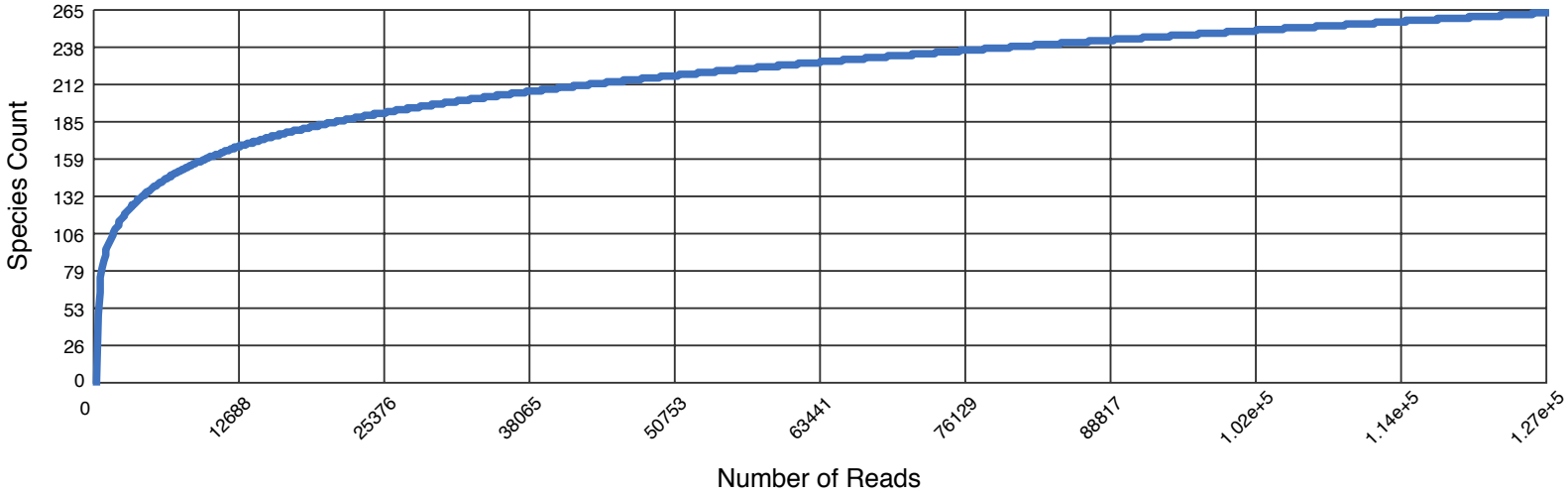
RAREFACTION CURVE HIDE

The plot below shows the rarefaction curve of annotated species richness. This curve is a plot of the total number of distinct species annotations as a function of the number of sequences sampled. On the left, a steep slope indicates that a large fraction of the species diversity remains to be discovered. If the curve becomes flatter to the right, a reasonable number of individuals is sampled: more intensive sampling is likely to yield only few additional species.

Sampling curves generally rise very quickly at first and then level off towards an asymptote as fewer new species are found per unit of individuals collected. These rarefaction curves are calculated from the table of species abundance. The curves represent the average number of different species annotations for subsamples of the the complete dataset.

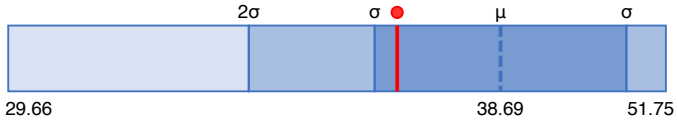
Download chart data

The image is currently dynamic. To be able to right-click/save the image, please click the static button static



ALPHA DIVERSITY [?] HIDE

α-Diversity = 30.590 species



The above image shows the range of α-diversity values in project sprague-april2015. The min, max, and mean values are shown, with the standard deviation ranges (σ

and 2σ) in different shades. The α-diversity of this metagenome is shown in red.

Alpha diversity summarizes the diversity of organisms in a sample with a single number. The alpha diversity of annotated samples can be estimated from the distribution of the species-level annotations.

Annotated species richness is the number of distinct species annotations in the combined MG-RAST dataset. Shannon diversity is an abundance-weighted average of the logarithm of the relative abundances of annotated species. The species-level annotations are from all the annotation source databases used by MG-RAST.

Download source data

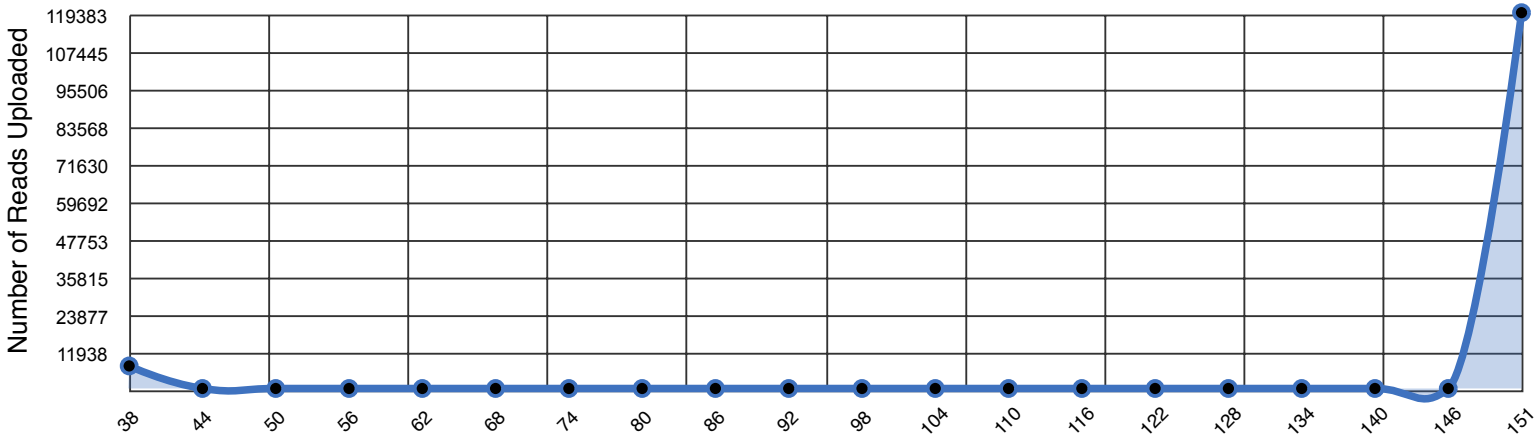
SEQUENCE LENGTH HISTOGRAM HIDE

The histograms below show the distribution of sequence lengths in basepairs for this metagenome. Each position represents the number of sequences within a length bp range.

The data used in these graphs are based on raw upload and post QC sequences.

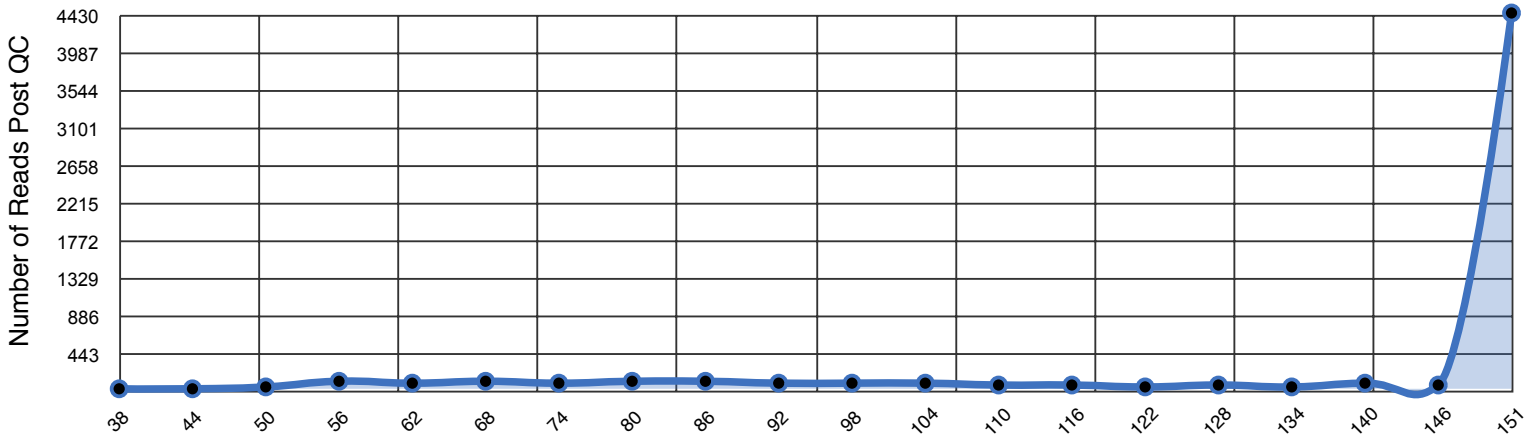
Download chart data

The image is currently dynamic. To be able to right-click/save the image, please click the static button static



Download chart data

The image is currently dynamic. To be able to right-click/save the image, please click the static button static

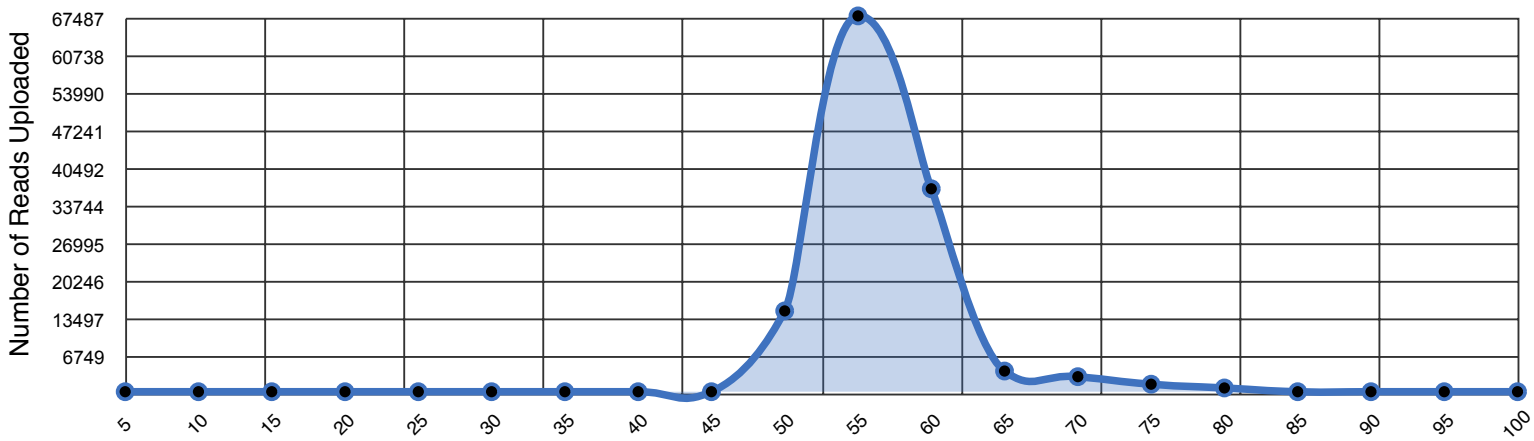


SEQUENCE GC DISTRIBUTION HIDE

The histograms below show the distribution of the GC percentage for this metagenome. Each position represents the number of sequences within a GC percentage range. The data used in these graphs is based on raw upload and post QC sequences.

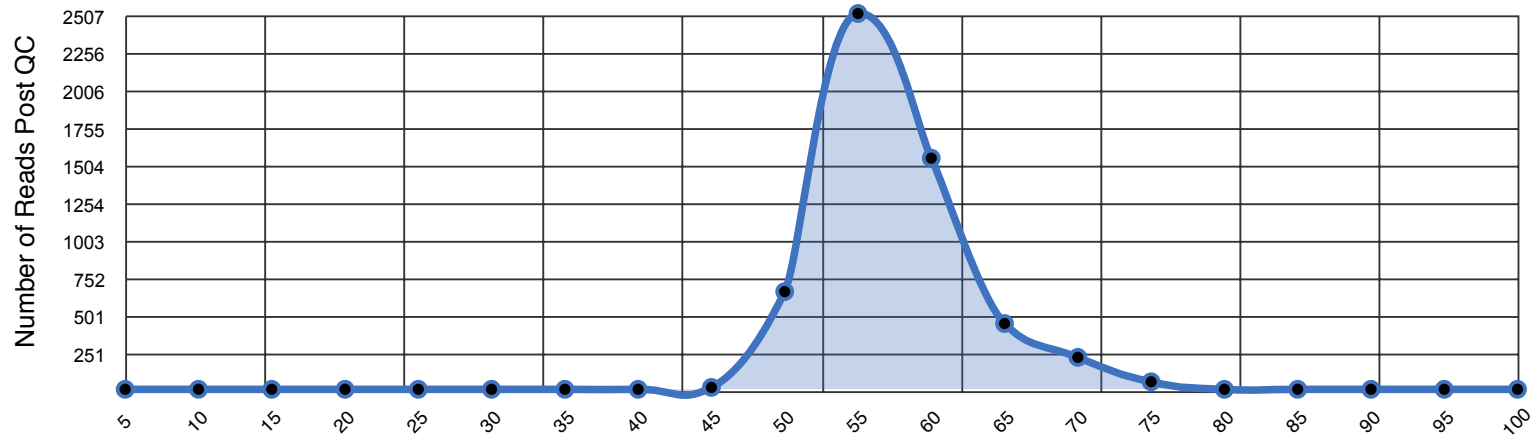
Download chart data

The image is currently dynamic. To be able to right-click/save the image, please click the static button static



Download chart data

The image is currently dynamic. To be able to right-click/save the image, please click the static button static



DOWNLOAD THIS METAGENOME

We provide download capabilities for the submitted sequences, metadata, and all files with results that are produced in the process of MG-RAST analysis on the download page for this metagenome. This includes fasta files with annotations using the M5NR.

We also provide access to the blat alignment summaries underlying our sequence analysis work on the download page.

Please note: The graphs on this page allow downloading the underlying information as tables. The search results and most of the pie-charts allow selecting the fraction of sequences in an element to work with in the workbench feature on the analysis page.

Analyze This Metagenome

The analysis page provides access to analysis and comparative tools including tables, bar charts, trees, principle coordinate analysis, heatmaps and various exports (including FASTA and QIIME). The workbench feature allows sub-selections of data to be used e.g. select all E. coli reads and then display the functional categories present just in E. coli reads across multiple data sets.

Search This Metagenome

Below searches return all predicted functions or organisms that contain the input text.

Search Functions

Search Organisms

Metadata [?] show